FITTING A STRAIGHT LINE WHEN BOTH VARIABLES ARE SUBJECT TO ERROR

M. S. Bartlett
University of Manchester, England

INTRODUCTION

A simple method of fitting a straight line when both variables are subject to error was examined by Wald (1) in 1940. The purpose of the present note is to present and illustrate a modification of Wald's method having the advantage in general of greater accuracy. Before any detailed exposition it will be as well to recall two important points:

(i) a distinction must be made between the linear regression equation of a variable y on a second variable x, and a linear functional relation between two variables Y and X masked by errors. The former equation is still available for prediction even if the variable x is subject to error, but is not necessarily appropriate for a functional relation when one exists.

(ii) it is possible to set up maximum likelihood equations for the second problem, but they do not lead to a unique solution without further assumptions, such as an assumption about the relative magnitude of the errors in x and y.

These points have been emphasized by many previous writers, for example, by Wald (1) or more recently by Lindley (2). In view of (ii) it is useful to consider, in the common case when the observations have equal weight, the following elementary method:

(a) For the location of the fitted straight line use as one point the mean coordinates $\bar{x}$, $\bar{y}$, just as in the least-squares method.

(b) For the slope, first divide the n plotted points into three groups, the equal numbers k in the two extreme groups being chosen to be as near $\frac{1}{3}n$ as possible (the three groups are non-overlapping when considered, say, in the x direction). The join of the mean coordinates $\bar{x}_1$, $\bar{y}_1$ and $\bar{x}_3$, $\bar{y}_3$ for the two extreme groups is used to determine the slope.

The only difference from Wald's original method is the use of three groups instead of two, for reasons which will be apparent from the results of the next section.[1] It will also be shown that Wald's confidence interval method of assessing the accuracy (under suitable conditions) may be adapted to the present method.

EFFICIENCY IN A SPECIAL CASE

To get some idea of the efficiency of the method its accuracy is determined in a special case where least-squares is appropriate. It is assumed that observations y are available for $n = 2l + 1$ values $x \equiv X$ not subject to error and spaced at equidistant unit intervals. The least-squares estimate is known to provide the linear combination of the y's providing an unbiased estimate of the true slope β in the functional relation

$$Y = \alpha + \beta X \tag{1}$$

with minimum variance when the differences $y - Y$ are uncorrelated and of constant variance σ^2. The least-squares estimate

$$b = \sum y(x - \bar{x}) / \sum (x - \bar{x})^2$$

has error variance $\sigma^2 / \sum (x - \bar{x})^2$, where $\sum (x - \bar{x})^2 = (\frac{1}{3})l(l + 1)(2l + 1)$ in the situation assumed in this section.

For comparison the error variance of the estimate

$$b' = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} \tag{2}$$

of the last section is easily evaluated for any value of k. It is given by

$$\frac{2\sigma^2}{k(\bar{x}_3 - \bar{x}_1)^2} = \frac{2\sigma^2}{k(2l - k + 1)^2} \cdot$$

The relative efficiency of b' is thus

$$E = \frac{3k(2l - k + 1)^2}{2l(l + 1)(2l + 1)} \cdot \tag{3}$$

This is a maximum when

$$(2l - k + 1)(2l + 1 - 3k) = 0$$

with relevant root $k = (\frac{1}{3})(2l + 1) = \frac{1}{3}n$.

[1] I am indebted to Professor Gerhard Tintner for drawing my attention to a previous discussion of this problem, with a similar conclusion, by Nair and Shrivastava (4)(see also Nair and Banerjee (5)). It might be noted that these authors propose using the two extreme groups out of three for location as well as slope, but recommendation (a) above is theoretically preferable. In the first of these two papers the extension of the method to fitting higher-order curves is also considered, though the optimum efficiency is not so high in such cases.

We then have

$$E = \frac{8(l + \frac{1}{2})^2}{9l(l + 1)} \geq \frac{8}{9}, \tag{4}$$

which may be compared with $E = (\frac{3}{4})(l + \frac{1}{2})^2/[l(l + 1)] \geq 3/4$ when $k = \frac{1}{2}n$. The higher efficiency of $k = \frac{1}{3}n$ compared with $k = \frac{1}{2}n$ suggests the adoption of $k = \frac{1}{3}n$ in preference to $k = \frac{1}{2}n$ in general. Indeed its high efficiency in the case examined above indicates the occasional value of the simple method proposed even in cases where the least-squares method is available.

ASSESSMENT OF ACCURACY IN THE GENERAL CASE

In the general problem it is assumed that both y and x are subject to error. To use Wald's confidence interval method it is assumed further that the n errors $\eta \equiv y - Y$ are independently and normally distributed with constant variance σ_η^2, similarly the n errors $\epsilon \equiv x - X$ are independent and normal with variance σ_ϵ^2; the x and y errors are moreover mutually independent, so that the variance of $\eta - \beta\epsilon$ is $\sigma_\eta^2 + \beta^2\sigma_\epsilon^2$.

Consider now possible 'estimates' of this last variance when β is known. If we write for the total sums of squares and products of x and y *within* the three groups

$$S_{xx} \equiv \sum\nolimits_1 (x - \bar{x}_1)^2 + \sum\nolimits_2 (x - \bar{x}_2)^2 + \sum\nolimits_3 (x - \bar{x}_3)^2$$

$$\begin{aligned} S_{xy} \equiv \sum\nolimits_1 (x - \bar{x}_1)(y - \bar{y}_1) &+ \sum\nolimits_2 (x - \bar{x}_2)(y - \bar{y}_2) \\ &+ \sum\nolimits_3 (x - \bar{x}_3)(y - \bar{y}_3) \end{aligned}$$

$$S_{yy} \equiv \sum\nolimits_1 (y - \bar{y}_1)^2 + \sum\nolimits_2 (y - \bar{y}_2)^2 + \sum\nolimits_3 (y - \bar{y}_3)^2,$$

where $\sum_i$ denotes summation over the observations in the i-th group, then $(S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx})/(n - 3)$ is an estimate of the variance $\sigma_\eta^2 + \beta^2\sigma_\epsilon^2$ with $n - 3$ degrees of freedom. The remaining 3 degrees of freedom are contained in the three group means. One is represented by the general mean, one by the difference between the means of the first and third groups to be used in the estimate of β; the third is represented by the difference between the mean of the second group and the general mean of the first and third groups.

For data with few observations it is advisable to make use of the last degree of freedom in the variance estimate, as in the numerical example considered later. Alternatively, if it is not so used, it remains available for testing the linearity of the true X, Y relation. In the former case, the appropriate square to be added to the numerator of the previous estimate is

$$\{(\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2)^2 - 2\beta(\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2)(\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2) + \beta^2(\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2)^2\}\left\{\frac{2}{k} + \frac{4}{n - k}\right\}$$

and the estimate $s^2(\beta)$ obtained with $n - 2$ now as the divisor will have $n - 2$ degrees of freedom.

Since

$$(\bar{x}_3 - \bar{x}_1)(b' - \beta) = (\bar{\eta}_3 - \beta\bar{\epsilon}_3) - (\bar{\eta}_1 - \beta\bar{\epsilon}_1),$$

when b' is given by (2), the left-hand quantity under the assumptions made in this section is normal with variance $(\sigma_\eta^2 + \beta^2\sigma_\epsilon^2)(2/k)$. This is subject to one qualification, that the errors in the x variable do not influence the allocation of the observations to the three groups. Such an effect may be neglected in many problems, particularly when the errors are small compared with the spacing of the observations at the points of division between the three groups; it will not be considered further here. A more detailed consideration of this point has been given by Wald (1).

Under the same assumptions we have

$$t = \frac{(\bar{x}_3 - \bar{x}_1)(b' - \beta)\sqrt{\tfrac{1}{2}k}}{s(\beta)} .$$

Although the denominator depends on β, this t-variate enables a confidence interval to be obtained for β. Thus for a value t corresponding to any chosen probability value we have the interval determined by the quadratic equation for β,

$$(5) \qquad (\bar{x}_3 - \bar{x}_1)^2(b' - \beta)^2\tfrac{1}{2}k = t^2(s_y^2 - 2\beta s_{xy} + \beta^2 s_x^2),$$

where $s^2(\beta) \equiv s_y^2 - 2\beta s_{xy} + \beta^2 s_x^2$.

If required, a similar method may be used to provide a joint confidence region for α and β. If $a \equiv \bar{y} - \beta\bar{x}$, then a is independent of the numerator of t and of $s(\beta)$, and hence

$$F = \frac{\tfrac{1}{2}\{n(a - \alpha)^2 + \tfrac{1}{2}k(\bar{x}_3 - \bar{x}_1)^2(b' - \beta)^2\}}{s^2(\beta)}$$

is a variance ratio with degrees of freedom 2, $n - 2$. For any chosen probability value the corresponding critical value of F will determine an ellipse as the boundary of the confidence region for α and β. This may be compared with the corresponding region for the least-squares method if it is known that $\sigma_\epsilon^2 = 0$; this region is similarly obtained from the variance ratio

$$F = \frac{\frac{1}{2}\{n(a - \alpha)^2 + (b - \beta)^2 \sum (x - \bar{x})^2\}}{s^2},$$

where s^2 is the usual variance estimate of $y - Y$ obtained from the residuals of y with $n - 2$ degrees of freedom.

If, as suggested earlier in this section, it is desired to examine the linearity of the functional relation, the variance estimate $s^2_{n-3}(\beta)$ of $\sigma^2_\eta + \beta^2\sigma^2_\epsilon$ with $n - 3$ degrees of freedom must be used. The further quantity

$$t = \frac{\{(\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2) - \beta(\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2)\}\left\{\frac{2}{k} + \frac{4}{n - k}\right\}^{-\frac{1}{2}}}{s_{n-3}(\beta)}$$

is then (if the linear relation is valid) also a t-variate with $n - 3$ degrees of freedom. It will be noticed that it involves the unknown slope β. When this is replaced by the estimate b', the resulting statistic is no longer exactly a t-variate, but might be treated approximately as such, especially when $\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2$ is small compared with $\bar{x}_3 - \bar{x}_1$.

NUMERICAL EXAMPLE

As a numerical example consider fitting a straight line to the data on penicillin 'assay' given by Davies (3, S 6.12). Six different concentrations of pure penicillin were set up on a plate on which an agar medium containing *B. subtilis* had been spread, and the mean circle diameters of the zones of inhibition of growth of the organisms were measured (for further details of the technique see S 5.41 of (3)). The concentration had negligible error, so that the standard least-squares method was available, the relation between circle diameter and log. concentration being linear. With circle diameter y in mms. and 1 penicillin unit per ml. as $x = 1$, and a two-fold increase in concentration as the unit for the x scale, the regression equation of y on x was

$$Y = 20.403 + 1.782(x - 3.5) = 14.166 + 1.782\,x \tag{6}$$

with a 95% confidence interval for the slope, based on the usual t-statistic, of (1.732, 1.832).

It is stressed that the data are considered again here purely in order to illustrate the present method. The six observations are divided into three groups:

y	15.87	17.78,	19.52	21.35,	23.13	24.77	(Total 122.42)
x	1	2 ,	3	4 ,	5	6	(Total 21)

$$b' = \frac{(24.77 + 23.13) - (17.78 + 15.87)}{(6 + 5) - (2 + 1)} = 1.781.$$

Hence the estimated relation is

$$Y = 20.403 + 1.781(X - 3.5) = 14.170 + 1.781X. \tag{7}$$

The sum of squares within each group has only one degree of freedom in this example, and may conveniently be calculated from the difference of the two observations per group. The other degree of freedom to be added is that for the contrast of the mean for the second group with the mean for the other two groups. This gives zero contribution for x, and for y

$$24.77 + 23.13 + 15.87 + 17.78 - 2(19.52 + 21.35) = -0.19$$

with appropriate divisor. Hence

$$4s_y^2 = \frac{(1.91)^2 + (1.83)^2 + (1.64)^2}{2} + \frac{(-0.19)^2}{12} = 4.8463$$

$$4s_{xy} = \frac{1 \times 1.91 + 1 \times 1.83 + 1 \times 1.64}{2} + \frac{(0 \times -0.19)}{12} = 2.69$$

$$4s_x^2 = \frac{1^2 + 1^2 + 1^2}{2} + \frac{0^2}{12} = 1.5.$$

Equation (5), with $t = 2.78$ for 4 degrees of freedom ($P = 0.05$), gives

$$16(1.781 - \beta)^2 = (2.78)^2(4.8463 - 2\beta \times 2.69 + 1.5\beta^2)/4$$

or

$$13.1018\beta^2 - 2\beta(23.2987) + 41.3879 = 0$$

or

$$\beta = 1.778 \pm 0.058.$$

Thus the 95% confidence interval for β by this method is (1.720, 1.836), an interval naturally slightly wider than the interval obtained by the least-squares method, since the assumption of no error in x has been dropped.

REFERENCES

(1) Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Stat.* 11, 284, 1940.

→ Lindley, D. V. Regression Lines and the Linear Functional Relationship. *J. Roy. Stat. Soc.* (*Suppl.*) 9, 218, 1947.

(3) Davies, O. L. (Editor). *Statistical Methods in Research and Production.* Oliver and Boyd, 1947.

(4) Nair, K. R. and Shrivastava, M. P. On a Simple Method of Curve Fitting. *Sankhyā* 6, 121, 1942.

(5) Nair, K. R. and Banerjee, K. S. A Note on Fitting of Straight Lines if Both Variables are Subject to Error. *Sankhyā* 6, 331, 1942.